Chapter 1

Set Theory and Logic

We adopt, as most mathematicians do, the naive point of view regarding set theory. We shall assume that what is meant by a *set* of objects is intuitively clear, and we shall proceed on that basis without analyzing the concept further. Such an analysis properly belongs to the foundations of mathematics and to mathematical logic, and it is not our purpose to initiate the study of those fields.

Logicians have analyzed set theory in great detail, and they have formulated axioms for the subject. Each of their axioms expresses a property of sets that mathematicians commonly accept, and collectively the axioms provide a foundation broad enough and strong enough that the rest of mathematics can be built on them.

It is unfortunately true that careless use of set theory, relying on intuition alone, can lead to contradictions. Indeed, one of the reasons for the axiomatization of set theory was to formulate rules for dealing with sets that would avoid these contradictions. Although we shall not deal with the axioms explicitly, the rules we follow in dealing with sets derive from them. In this book, you will learn how to deal with sets in an "apprentice" fashion, by observing how we handle them and by working with them yourself. At some point of your studies, you may wish to study set theory more carefully and in greater detail; then a course in logic or foundations will be in order.

§1 Fundamental Concepts

Here we introduce the ideas of set theory, and establish the basic terminology and notation. We also discuss some points of elementary logic that, in our experience, are apt to cause confusion.

Basic Notation

Commonly we shall use capital letters A, B, ... to denote sets, and lowercase letters a, b, ... to denote the ***objects*** or ***elements*** belonging to these sets. If an object a belongs to a set A, we express this fact by the notation

$$a \in A.$$

If a does not belong to A, we express this fact by writing

$$a \notin A.$$

The equality symbol $=$ is used throughout this book to mean *logical identity*. Thus, when we write $a = b$, we mean that "a" and "b" are symbols for the same object. This is what one means in arithmetic, for example, when one writes $\frac{2}{4} = \frac{1}{2}$. Similarly, the equation $A = B$ states that "A" and "B" are symbols for the same set; that is, A and B consist of precisely the same objects.

If a and b are different objects, we write $a \neq b$; and if A and B are different sets, we write $A \neq B$. For example, if A is the set of all nonnegative real numbers, and B is the set of all positive real numbers, then $A \neq B$, because the number 0 belongs to A and not to B.

We say that A is a ***subset*** of B if every element of A is also an element of B; and we express this fact by writing

$$A \subset B.$$

Nothing in this definition requires A to be different from B; in fact, if $A = B$, it is true that both $A \subset B$ and $B \subset A$. If $A \subset B$ and A is different from B, we say that A is a ***proper subset*** of B, and we write

$$A \subsetneq B.$$

The relations $\subset$ and $\subsetneq$ are called ***inclusion*** and ***proper inclusion***, respectively. If $A \subset B$, we also write $B \supset A$, which is read "B ***contains*** A."

How does one go about specifying a set? If the set has only a few elements, one can simply list the objects in the set, writing "A is the set consisting of the elements a, b, and c." In symbols, this statement becomes

$$A = \{a, b, c\},$$

where braces are used to enclose the list of elements.

The usual way to specify a set, however, is to take some set A of objects and some *property* that elements of A may or may not possess, and to form the set consisting of all elements of A having that property. For instance, one might take the set of real numbers and form the subset B consisting of all even integers. In symbols, this statement becomes

$$B = \{x \mid x \text{ is an even integer}\}.$$

Here the braces stand for the words "the set of," and the vertical bar stands for the words "such that." The equation is read "B is the set of all x such that x is an even integer."

The Union of Sets and the Meaning of "or"

Given two sets A and B, one can form a set from them that consists of all the elements of A together with all the elements of B. This set is called the ***union*** of A and B and is denoted by $A \cup B$. Formally, we define

$$A \cup B = \{x \mid x \in A \text{ or } x \in B\}.$$

But we must pause at this point and make sure exactly what we mean by the statement "$x \in A$ or $x \in B$."

In ordinary everyday English, the word "or" is ambiguous. Sometimes the statement "P or Q" means "P or Q, or both" and sometimes it means "P or Q, but not both." Usually one decides from the context which meaning is intended. For example, suppose I spoke to two students as follows:

> "Miss Smith, every student registered for this course has taken either a course in linear algebra or a course in analysis."
>
> "Mr. Jones, either you get a grade of at least 70 on the final exam or you will flunk this course."

In the context, Miss Smith knows perfectly well that I mean "everyone has had linear algebra or analysis, or both," and Mr. Jones knows I mean "either he gets at least 70 or he flunks, but not both." Indeed, Mr. Jones would be exceedingly unhappy if both statements turned out to be true!

In mathematics, one cannot tolerate such ambiguity. One has to pick just one meaning and stick with it, or confusion will reign. Accordingly, mathematicians have agreed that they will use the word "or" in the first sense, so that the statement "P or Q" always means "P or Q, or both." If one means "P or Q, but not both," then one has to include the phrase "but not both" explicitly.

With this understanding, the equation defining $A \cup B$ is unambiguous; it states that $A \cup B$ is the set consisting of all elements x that belong to A or to B or to both.

The Intersection of Sets, the Empty Set, and the Meaning of "If . . . Then"

Given sets A and B, another way one can form a set is to take the common part of A and B. This set is called the ***intersection*** of A and B and is denoted by $A \cap B$. Formally, we define

$$A \cap B = \{x \mid x \in A \text{ and } x \in B\}.$$

But just as with the definition of $A \cup B$, there is a difficulty. The difficulty is not in the meaning of the word "and"; it is of a different sort. It arises when the sets A and B happen to have no elements in common. What meaning does the symbol $A \cap B$ have in such a case?

To take care of this eventuality, we make a special convention. We introduce a special set that we call the ***empty set***, denoted by $\varnothing$, which we think of as "the set having no elements."

Using this convention, we express the statement that A and B have no elements in common by the equation

$$A \cap B = \varnothing.$$

We also express this fact by saying that A and B are ***disjoint***.

Now some students are bothered by the notion of an "empty set." "How," they say, "can you have a set with nothing in it?" The problem is similar to that which arose many years ago when the number 0 was first introduced.

The empty set is only a convention, and mathematics could very well get along without it. But it is a very convenient convention, for it saves us a good deal of awkwardness in stating theorems and in proving them. Without this convention, for instance, one would have to prove that the two sets A and B do have elements in common before one could use the notation $A \cap B$. Similarly, the notation

$$C = \{x \mid x \in A \text{ and } x \text{ has a certain property}\}$$

could not be used if it happened that no element x of A had the given property. It is much more convenient to agree that $A \cap B$ and C equal the empty set in such cases.

Since the empty set $\varnothing$ is merely a convention, we must make conventions relating it to the concepts already introduced. Because $\varnothing$ is thought of as "the set with no elements," it is clear we should make the convention that for each object x, the relation $x \in \varnothing$ does not hold. Similarly, the definitions of union and intersection show that for every set A we should have the equations

$$A \cup \varnothing = A \quad \text{and} \quad A \cap \varnothing = \varnothing.$$

The inclusion relation is a bit more tricky. Given a set A, should we agree that $\varnothing \subset A$? Once more, we must be careful about the way mathematicians use the English language. The expression $\varnothing \subset A$ is a shorthand way of writing the sentence, "Every element that belongs to the empty set also belongs to the set A." Or to put it more

formally, "For every object x, if x belongs to the empty set, then x also belongs to the set A."

Is this statement true or not? Some might say "yes" and others say "no." You will never settle the question by argument, only by agreement. This is a statement of the form "If P, then Q," and in everyday English the meaning of the "if ... then" construction is ambiguous. It always means that if P is true, then Q is true also. Sometimes that is all it means; other times it means something more: that if P is false, Q must be false. Usually one decides from the context which interpretation is correct.

The situation is similar to the ambiguity in the use of the word "or." One can reformulate the examples involving Miss Smith and Mr. Jones to illustrate the ambiguity. Suppose I said the following:

> "Miss Smith, if any student registered for this course has not taken a course in linear algebra, then he has taken a course in analysis."
>
> "Mr. Jones, if you get a grade below 70 on the final, you are going to flunk this course."

In the context, Miss Smith understands that if a student in the course has not had linear algebra, then he has taken analysis, but if he has had linear algebra, he may or may not have taken analysis as well. And Mr. Jones knows that if he gets a grade below 70, he will flunk the course, but if he gets a grade of at least 70, he will pass.

Again, mathematics cannot tolerate ambiguity, so a choice of meanings must be made. Mathematicians have agreed always to use "if ... then" in the first sense, so that a statement of the form "If P, then Q" means that if P is true, Q is true also, but if P is false, Q may be either true or false.

As an example, consider the following statement about real numbers:

$$\textit{If } x > 0, \textit{ then } x^3 \neq 0.$$

It is a statement of the form, "If P, then Q," where P is the phrase "$x > 0$" (called the ***hypothesis*** of the statement) and Q is the phrase "$x^3 \neq 0$" (called the ***conclusion*** of the statement). This is a true statement, for in every case for which the hypothesis $x > 0$ holds, the conclusion $x^3 \neq 0$ holds as well.

Another true statement about real numbers is the following:

$$\textit{If } x^2 < 0, \textit{ then } x = 23;$$

in every case for which the hypothesis holds, the conclusion holds as well. Of course, it happens in this example that there are no cases for which the hypothesis holds. A statement of this sort is sometimes said to be ***vacuously true***.

To return now to the empty set and inclusion, we see that the inclusion $\varnothing \subset A$ does hold for every set A. Writing $\varnothing \subset A$ is the same as saying, "If $x \in \varnothing$, then $x \in A$," and this statement is vacuously true.

Contrapositive and Converse

Our discussion of the "if ... then" construction leads us to consider another point of elementary logic that sometimes causes difficulty. It concerns the relation between a *statement*, its *contrapositive*, and its *converse*.

Given a statement of the form "If P, then Q," its ***contrapositive*** is defined to be the statement "If Q is not true, then P is not true." For example, the contrapositive of the statement

$$\textit{If } x > 0, \textit{ then } x^3 \neq 0,$$

is the statement

$$\textit{If } x^3 = 0, \textit{ then it is not true that } x > 0.$$

Note that both the statement and its contrapositive are true. Similarly, the statement

$$\textit{If } x^2 < 0, \textit{ then } x = 23,$$

has as its contrapositive the statement

$$\textit{If } x \neq 23, \textit{ then it is not true that } x^2 < 0.$$

Again, both are true statements about real numbers.

These examples may make you suspect that there is some relation between a statement and its contrapositive. And indeed there is; they are two ways of saying precisely the same thing. Each is true if and only if the other is true; they are *logically equivalent*.

This fact is not hard to demonstrate. Let us introduce some notation first. As a shorthand for the statement "If P, then Q," we write

$$P \Longrightarrow Q,$$

which is read "P implies Q." The contrapositive can then be expressed in the form

$$(\text{not } Q) \Longrightarrow (\text{not } P),$$

where "not Q" stands for the phrase "Q is not true."

Now the only way in which the statement "$P \Rightarrow Q$" can fail to be correct is if the hypothesis P is true and the conclusion Q is false. Otherwise it is correct. Similarly, the only way in which the statement (not Q) $\Rightarrow$ (not P) can fail to be correct is if the hypothesis "not Q" is true and the conclusion "not P" is false. This is the same as saying that Q is false and P is true. And this, in turn, is precisely the situation in which $P \Rightarrow Q$ fails to be correct. Thus, we see that the two statements are either both correct or both incorrect; they are logically equivalent. Therefore, we shall accept a proof of the statement "not $Q \Rightarrow$ not P" as a proof of the statement "$P \Rightarrow Q$."

There is another statement that can be formed from the statement $P \Rightarrow Q$. It is the statement

$$Q \Longrightarrow P,$$

which is called the ***converse*** of $P \Rightarrow Q$. One must be careful to distinguish between a statement's converse and its contrapositive. Whereas a statement and its contrapositive are logically equivalent, the truth of a statement says nothing at all about the truth or falsity of its converse. For example, the true statement

$$\textit{If } x > 0, \textit{ then } x^3 \neq 0,$$

has as its converse the statement

$$\textit{If } x^3 \neq 0, \textit{ then } x > 0,$$

which is false. Similarly, the true statement

$$\textit{If } x^2 < 0, \textit{ then } x = 23,$$

has as its converse the statement

$$\textit{If } x = 23, \textit{ then } x^2 < 0,$$

which is false.

If it should happen that both the statement $P \Rightarrow Q$ and its converse $Q \Rightarrow P$ are true, we express this fact by the notation

$$P \Longleftrightarrow Q,$$

which is read "P holds if and only if Q holds."

Negation

If one wishes to form the contrapositive of the statement $P \Rightarrow Q$, one has to know how to form the statement "not P," which is called the ***negation*** of P. In many cases, this causes no difficulty; but sometimes confusion occurs with statements involving the phrases "for every" and "for at least one." These phrases are called *logical quantifiers*.

To illustrate, suppose that X is a set, A is a subset of X, and P is a statement about the general element of X. Consider the following statement:

$$(*) \qquad \textit{For every } x \in A, \textit{ statement } P \textit{ holds.}$$

How does one form the negation of this statement? Let us translate the problem into the language of sets. Suppose that we let B denote the set of all those elements x of X for which P holds. Then statement $(*)$ is just the statement that A is a subset of B. What is its negation? Obviously, the statement that A is *not* a subset of B; that is, the statement that there exists at least one element of A that does not belong to B. Translating back into ordinary language, this becomes

$$\textit{For at least one } x \in A, \textit{ statement } P \textit{ does not hold.}$$

Therefore, to form the negation of statement $(*)$, one replaces the quantifier "for every" by the quantifier "for at least one," and one replaces statement P by *its* negation.

The process works in reverse just as well; the negation of the statement

For at least one $x \in A$, statement Q holds,

is the statement

For every $x \in A$, statement Q does not hold.

The Difference of Two Sets

We return now to our discussion of sets. There is one other operation on sets that is occasionally useful. It is the ***difference*** of two sets, denoted by $A - B$, and defined as the set consisting of those elements of A that are not in B. Formally,

$$A - B = \{x \mid x \in A \text{ and } x \notin B\}.$$

It is sometimes called the ***complement*** of B relative to A, or the complement of B *in* A.

Our three set operations are represented schematically in Figure 1.1.

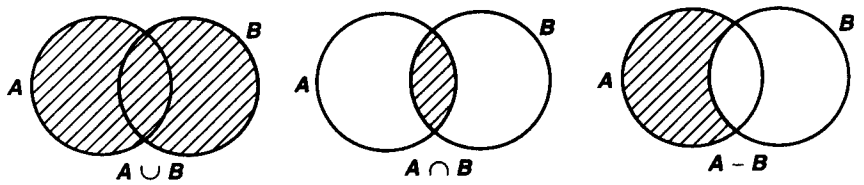

Figure 1.1

Rules of Set Theory

Given several sets, one may form new sets by applying the set-theoretic operations to them. As in algebra, one uses parentheses to indicate in what order the operations are to be performed. For example, $A \cup (B \cap C)$ denotes the union of the two sets A and $B \cap C$, while $(A \cup B) \cap C$ denotes the intersection of the two sets $A \cup B$ and C. The sets thus formed are quite different, as Figure 1.2 shows.

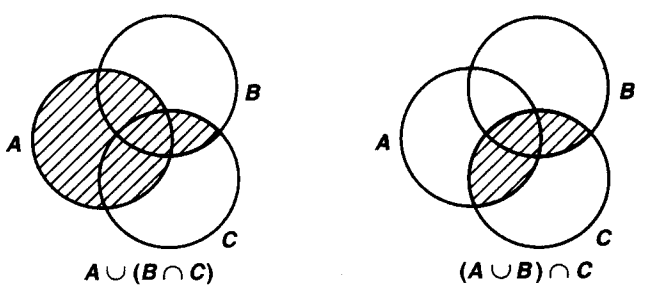

Figure 1.2

Sometimes different combinations of operations lead to the same set; when that happens, one has a rule of set theory. For instance, it is true that for any sets A, B, and C the equation

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

holds. The equation is illustrated in Figure 1.3; the shaded region represents the set in question, as you can check mentally. This equation can be thought of as a "distributive law" for the operations $\cap$ and $\cup$.

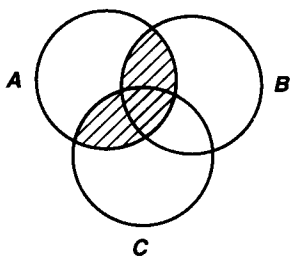

Figure 1.3

Other examples of set-theoretic rules include the second "distributive law,"

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C),$$

and *DeMorgan's laws*,

$$A - (B \cup C) = (A - B) \cap (A - C),$$
$$A - (B \cap C) = (A - B) \cup (A - C).$$

We leave it to you to check these rules. One can state other rules of set theory, but these are the most important ones. DeMorgan's laws are easier to remember if you verbalize them as follows:

> *The complement of the union equals the intersection of the complements.*
> *The complement of the intersection equals the union of the complements.*

Collections of Sets

The objects belonging to a set may be of any sort. One can consider the set of all even integers, and the set of all blue-eyed people in Nebraska, and the set of all decks of playing cards in the world. Some of these are of limited mathematical interest, we admit! But the third example illustrates a point we have not yet mentioned: namely, that the objects belonging to a set may *themselves* be sets. For a deck of cards is itself a set, one consisting of pieces of pasteboard with certain standard designs printed on them. The set of all decks of cards in the world is thus a set whose elements are themselves sets (of pieces of pasteboard).

We now have another way to form new sets from old ones. Given a set A, we can consider sets whose elements are subsets of A. In particular, we can consider the set of all subsets of A. This set is sometimes denoted by the symbol $\mathcal{P}(A)$ and is called the ***power set*** of A (for reasons to be explained later).

When we have a set whose elements are sets, we shall often refer to it as a ***collection*** of sets and denote it by a script letter such as $\mathcal{A}$ or $\mathcal{B}$. This device will help us in keeping things straight in arguments where we have to consider objects, and sets of objects, and collections of sets of objects, all at the same time. For example, we might use $\mathcal{A}$ to denote the collection of all decks of cards in the world, letting an ordinary capital letter A denote a deck of cards and a lowercase letter a denote a single playing card.

A certain amount of care with notation is needed at this point. We make a distinction between the object a, which is an *element* of a set A, and the one-element set $\{a\}$, which is a *subset* of A. To illustrate, if A is the set $\{a, b, c\}$, then the statements

$$a \in A, \qquad \{a\} \subset A, \qquad \text{and} \qquad \{a\} \in \mathcal{P}(A)$$

are all correct, but the statements $\{a\} \in A$ and $a \subset A$ are not.

Arbitrary Unions and Intersections

We have already defined what we mean by the union and the intersection of two sets. There is no reason to limit ourselves to just two sets, for we can just as well form the union and intersection of arbitrarily many sets.

Given a collection $\mathcal{A}$ of sets, the ***union*** of the elements of $\mathcal{A}$ is defined by the equation

$$\bigcup_{A \in \mathcal{A}} A = \{x \mid x \in A \text{ for at least one } A \in \mathcal{A}\}.$$

The ***intersection*** of the elements of $\mathcal{A}$ is defined by the equation

$$\bigcap_{A \in \mathcal{A}} A = \{x \mid x \in A \text{ for every } A \in \mathcal{A}\}.$$

There is no problem with these definitions if one of the elements of $\mathcal{A}$ happens to be the empty set. But it is a bit tricky to decide what (if anything) these definitions mean if we allow $\mathcal{A}$ to be the empty collection. Applying the definitions literally, we see that no element x satisfies the defining property for the union of the elements of $\mathcal{A}$. So it is reasonable to say that

$$\bigcup_{A \in \mathcal{A}} A = \varnothing$$

if $\mathcal{A}$ is empty. On the other hand, every x satisfies (vacuously) the defining property for the intersection of the elements of $\mathcal{A}$. The question is, every x in what set? If one has a given large set X that is specified at the outset of the discussion to be one's "universe of discourse," and one considers only subsets of X throughout, it is reasonable to let

$$\bigcap_{A \in \mathcal{A}} A = X$$

when $\mathcal{A}$ is empty. Not all mathematicians follow this convention, however. To avoid difficulty, *we shall not define the intersection when $\mathcal{A}$ is empty.*

Cartesian Products

There is yet another way of forming new sets from old ones; it involves the notion of an "ordered pair" of objects. When you studied analytic geometry, the first thing you did was to convince yourself that after one has chosen an x-axis and a y-axis in the plane, every point in the plane can be made to correspond to a unique ordered pair (x, y) of real numbers. (In a more sophisticated treatment of geometry, the plane is more likely to be *defined* as the set of all ordered pairs of real numbers!)

The notion of ordered pair carries over to general sets. Given sets A and B, we define their cartesian product $A \times B$ to be the set of all ordered pairs (a, b) for which a is an element of A and b is an element of B. Formally,

$$A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}.$$

This definition assumes that the concept of "ordered pair" is already given. It can be taken as a primitive concept, as was the notion of "set"; or it can be given a definition in terms of the set operations already introduced. One definition in terms of set operations is expressed by the equation

$$(a, b) = \{\{a\}, \{a, b\}\};$$

it defines the ordered pair (a, b) as a collection of sets. If $a \neq b$, this definition says that (a, b) is a collection containing two sets, one of which is a one-element set and the other a two-element set. The *first coordinate* of the ordered pair is defined to be the element belonging to both sets, and the *second coordinate* is the element belonging to only one of the sets. If $a = b$, then (a, b) is a collection containing only one set $\{a\}$, since $\{a, b\} = \{a, a\} = \{a\}$ in this case. Its first coordinate and second coordinate both equal the element in this single set.

I think it is fair to say that most mathematicians think of an ordered pair as a primitive concept rather than thinking of it as a collection of sets!

Let us make a comment on notation. It is an unfortunate fact that the notation (a, b) is firmly established in mathematics with two entirely different meanings. One meaning, as an ordered pair of objects, we have just discussed. The other meaning is the one you are familiar with from analysis; if a and b are real numbers, the symbol (a, b) is used to denote the interval consisting of all numbers x such that $a < x < b$. Most of the time, this conflict in notation will cause no difficulty because the meaning will be clear from the context. Whenever a situation occurs where confusion is possible, we shall adopt a different notation for the ordered pair (a, b), denoting it by the symbol

$$a \times b$$

instead.

Exercises

1. Check the distributive laws for $\cup$ and $\cap$ and DeMorgan's laws.
2. Determine which of the following statements are true for all sets A, B, C, and D. If a double implication fails, determine whether one or the other of the possible implications holds. If an equality fails, determine whether the statement becomes true if the "equals" symbol is replaced by one or the other of the inclusion symbols $\subset$ or $\supset$.
   (a) $A \subset B$ and $A \subset C \Leftrightarrow A \subset (B \cup C)$.
   (b) $A \subset B$ or $A \subset C \Leftrightarrow A \subset (B \cup C)$.
   (c) $A \subset B$ and $A \subset C \Leftrightarrow A \subset (B \cap C)$.
   (d) $A \subset B$ or $A \subset C \Leftrightarrow A \subset (B \cap C)$.
   (e) $A - (A - B) = B$.
   (f) $A - (B - A) = A - B$.
   (g) $A \cap (B - C) = (A \cap B) - (A \cap C)$.
   (h) $A \cup (B - C) = (A \cup B) - (A \cup C)$.
   (i) $(A \cap B) \cup (A - B) = A$.
   (j) $A \subset C$ and $B \subset D \Rightarrow (A \times B) \subset (C \times D)$.
   (k) The converse of (j).
   (l) The converse of (j), assuming that A and B are nonempty.
   (m) $(A \times B) \cup (C \times D) = (A \cup C) \times (B \cup D)$.
   (n) $(A \times B) \cap (C \times D) = (A \cap C) \times (B \cap D)$.
   (o) $A \times (B - C) = (A \times B) - (A \times C)$.
   (p) $(A - B) \times (C - D) = (A \times C - B \times C) - A \times D$.
   (q) $(A \times B) - (C \times D) = (A - C) \times (B - D)$.
3. (a) Write the contrapositive and converse of the following statement: "If $x < 0$, then $x^2 - x > 0$," and determine which (if any) of the three statements are true.
   (b) Do the same for the statement "If $x > 0$, then $x^2 - x > 0$."
4. Let A and B be sets of real numbers. Write the negation of each of the following statements:
   (a) For every $a \in A$, it is true that $a^2 \in B$.
   (b) For at least one $a \in A$, it is true that $a^2 \in B$.
   (c) For every $a \in A$, it is true that $a^2 \notin B$.
   (d) For at least one $a \notin A$, it is true that $a^2 \in B$.
5. Let $\mathcal{A}$ be a nonempty collection of sets. Determine the truth of each of the following statements and of their converses:
   (a) $x \in \bigcup_{A \in \mathcal{A}} A \Rightarrow x \in A$ for at least one $A \in \mathcal{A}$.
   (b) $x \in \bigcup_{A \in \mathcal{A}} A \Rightarrow x \in A$ for every $A \in \mathcal{A}$.
   (c) $x \in \bigcap_{A \in \mathcal{A}} A \Rightarrow x \in A$ for at least one $A \in \mathcal{A}$.
   (d) $x \in \bigcap_{A \in \mathcal{A}} A \Rightarrow x \in A$ for every $A \in \mathcal{A}$.
6. Write the contrapositive of each of the statements of Exercise 5.

7. Given sets A, B, and C, express each of the following sets in terms of A, B, and C, using the symbols $\cup$, $\cap$, and $-$.

$$D = \{x \mid x \in A \text{ and } (x \in B \text{ or } x \in C)\},$$
$$E = \{x \mid (x \in A \text{ and } x \in B) \text{ or } x \in C\},$$
$$F = \{x \mid x \in A \text{ and } (x \in B \Rightarrow x \in C)\}.$$

8. If a set A has two elements, show that $\mathcal{P}(A)$ has four elements. How many elements does $\mathcal{P}(A)$ have if A has one element? Three elements? No elements? Why is $\mathcal{P}(A)$ called the power set of A?
9. Formulate and prove DeMorgan's laws for arbitrary unions and intersections.
10. Let $\mathbb{R}$ denote the set of real numbers. For each of the following subsets of $\mathbb{R} \times \mathbb{R}$, determine whether it is equal to the cartesian product of two subsets of $\mathbb{R}$.
    (a) $\{(x, y) \mid x \text{ is an integer}\}$.
    (b) $\{(x, y) \mid 0 < y \leq 1\}$.
    (c) $\{(x, y) \mid y > x\}$.
    (d) $\{(x, y) \mid x \text{ is not an integer and } y \text{ is an integer}\}$.
    (e) $\{(x, y) \mid x^2 + y^2 < 1\}$.

§2 Functions

The concept of *function* is one you have seen many times already, so it is hardly necessary to remind you how central it is to all mathematics. In this section, we give the precise mathematical definition, and we explore some of the associated concepts.

A function is usually thought of as a *rule* that assigns to each element of a set A, an element of a set B. In calculus, a function is often given by a simple formula such as $f(x) = 3x^2 + 2$ or perhaps by a more complicated formula such as

$$f(x) = \sum_{k=1}^{\infty} x^k.$$

One often does not even mention the sets A and B explicitly, agreeing to take A to be the set of all real numbers for which the rule makes sense and B to be the set of all real numbers.

As one goes further in mathematics, however, one needs to be more precise about what a function is. Mathematicians *think* of functions in the way we just described, but the definition they use is more exact. First, we define the following:

Definition. A ***rule of assignment*** is a subset r of the cartesian product $C \times D$ of two sets, having the property that each element of C appears as the first coordinate of *at most one* ordered pair belonging to r.

Thus, a subset r of $C \times D$ is a rule of assignment if

$$[(c, d) \in r \text{ and } (c, d') \in r] \Longrightarrow [d = d'].$$

We think of r as a way of assigning, to the element c of C, the element d of D for which $(c, d) \in r$.

Given a rule of assignment r, the ***domain*** of r is defined to be the subset of C consisting of all first coordinates of elements of r, and the ***image set*** of r is defined as the subset of D consisting of all second coordinates of elements of r. Formally,

$$\text{domain } r = \{c \mid \text{there exists } d \in D \text{ such that } (c, d) \in r\},$$
$$\text{image } r = \{d \mid \text{there exists } c \in C \text{ such that } (c, d) \in r\}.$$

Note that given a rule of assignment r, its domain and image are entirely determined.

Now we can say what a function is.

Definition. A ***function*** f is a rule of assignment r, together with a set B that contains the image set of r. The domain A of the rule r is also called the ***domain*** of the function f; the image set of r is also called the ***image set*** of f; and the set B is called the ***range*** of f.†

If f is a function having domain A and range B, we express this fact by writing

$$f : A \longrightarrow B,$$

which is read "f is a function from A to B," or "f is a mapping from A into B," or simply "f maps A into B." One sometimes visualizes f as a geometric transformation physically carrying the points of A to points of B.

If $f : A \to B$ and if a is an element of A, we denote by $f(a)$ the unique element of B that the rule determining f assigns to a; it is called the ***value*** of f at a, or sometimes the ***image*** of a under f. Formally, if r is the rule of the function f, then $f(a)$ denotes the unique element of B such that $(a, f(a)) \in r$.

Using this notation, one can go back to defining functions almost as one did before, with no lack of rigor. For instance, one can write (letting $\mathbb{R}$ denote the real numbers)

> "Let f be the function whose rule is $\{(x, x^3 + 1) \mid x \in \mathbb{R}\}$ and whose range is $\mathbb{R}$,"

or one can equally well write

> "Let $f : \mathbb{R} \to \mathbb{R}$ be the function such that $f(x) = x^3 + 1$."

Both sentences specify precisely the same function. But the sentence "Let f be the function $f(x) = x^3 + 1$" is no longer adequate for specifying a function because it specifies neither the domain nor the range of f.

†Analysts are apt to use the word "range" to denote what we have called the "image set" of f. They avoid giving the set B a name.

Definition. If $f : A \to B$ and if A_0 is a subset of A, we define the ***restriction*** of f to A_0 to be the function mapping A_0 into B whose rule is

$$\{(a, f(a)) \mid a \in A_0\}.$$

It is denoted by $f|A_0$, which is read "f restricted to A_0."

EXAMPLE 1. Let $\mathbb{R}$ denote the real numbers and let $\bar{\mathbb{R}}_+$ denote the nonnegative reals. Consider the functions

$$\begin{array}{llll} f : \mathbb{R} \longrightarrow \mathbb{R} & \text{defined by} & f(x) = x^2, \\ g : \bar{\mathbb{R}}_+ \longrightarrow \mathbb{R} & \text{defined by} & g(x) = x^2, \\ h : \mathbb{R} \longrightarrow \bar{\mathbb{R}}_+ & \text{defined by} & h(x) = x^2, \\ k : \bar{\mathbb{R}}_+ \longrightarrow \bar{\mathbb{R}}_+ & \text{defined by} & k(x) = x^2. \end{array}$$

The function g is different from the function f because their rules are different subsets of $\mathbb{R} \times \mathbb{R}$; it is the restriction of f to the set $\bar{\mathbb{R}}_+$. The function h is also different from f, even though their rules are the same set, because the range specified for h is different from the range specified for f. The function k is different from all of these. These functions are pictured in Figure 2.1.

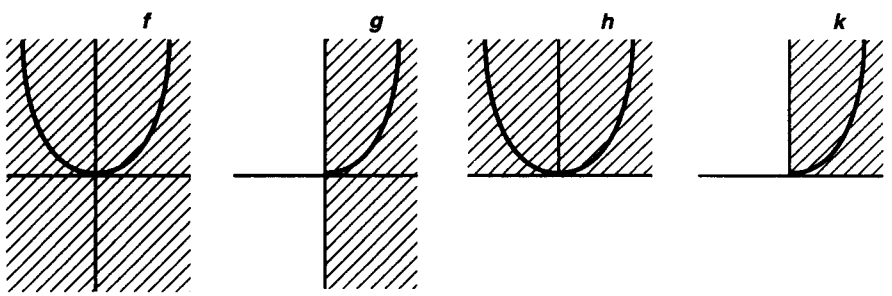

Figure 2.1

Restricting the domain of a function and changing its range are two ways of forming a new function from an old one. Another way is to form the composite of two functions.

Definition. Given functions $f : A \to B$ and $g : B \to C$, we define the ***composite*** $g \circ f$ of f and g as the function $g \circ f : A \to C$ defined by the equation $(g \circ f)(a) = g(f(a))$.

Formally, $g \circ f : A \to C$ is the function whose rule is

$$\{(a, c) \mid \text{For some } b \in B,\ f(a) = b \text{ and } g(b) = c\}.$$

We often picture the composite $g \circ f$ as involving a physical movement of the point a to the point $f(a)$, and then to the point $g(f(a))$, as illustrated in Figure 2.2.

Note that $g \circ f$ is defined only when the range of f *equals* the domain of g.

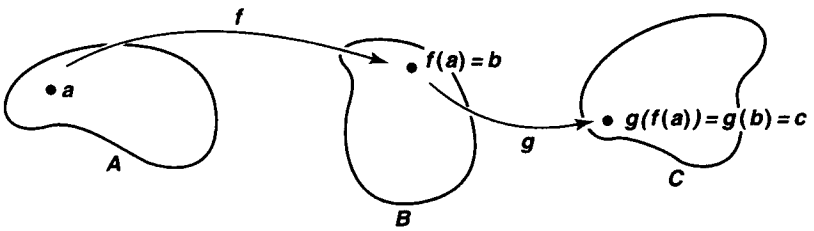

Figure 2.2

EXAMPLE 2. The composite of the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = 3x^2 + 2$ and the function $g : \mathbb{R} \to \mathbb{R}$ given by $g(x) = 5x$ is the function $g \circ f : \mathbb{R} \to \mathbb{R}$ given by

$$(g \circ f)(x) = g(f(x)) = g(3x^2 + 2) = 5(3x^2 + 2).$$

The composite $f \circ g$ can also be formed in this case; it is the quite different function $f \circ g : \mathbb{R} \to \mathbb{R}$ given by

$$(f \circ g)(x) = f(g(x)) = f(5x) = 3(5x)^2 + 2.$$

Definition. A function $f : A \to B$ is said to be ***injective*** (or ***one-to-one***) if for each pair of distinct points of A, their images under f are distinct. It is said to be ***surjective*** (or f is said to map A ***onto*** B) if every element of B is the image of some element of A under the function f. If f is both injective and surjective, it is said to be ***bijective*** (or is called a ***one-to-one correspondence***).

More formally, f is injective if

$$[f(a) = f(a')] \Longrightarrow [a = a'],$$

and f is surjective if

$$[b \in B] \Longrightarrow [b = f(a) \text{ for at least one } a \in A].$$

Injectivity of f depends only on the rule of f; surjectivity depends on the range of f as well. You can check that the composite of two injective functions is injective, and the composite of two surjective functions is surjective; it follows that the composite of two bijective functions is bijective.

If f is bijective, there exists a function from B to A called the ***inverse*** of f. It is denoted by f^{-1} and is defined by letting $f^{-1}(b)$ be that unique element a of A for which $f(a) = b$. Given $b \in B$, the fact that f is surjective implies that there *exists* such an element $a \in A$; the fact that f is injective implies that there is *only one* such element a. It is easy to see that if f is bijective, f^{-1} is also bijective.

EXAMPLE 3. Consider again the functions f, g, h, and k of Figure 2.1. The function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$ is neither injective nor surjective. Its restriction g to the nonnegative reals is injective but not surjective. The function $h : \mathbb{R} \to \bar{\mathbb{R}}_+$ obtained from f

by changing the range is surjective but not injective. The function $k : \tilde{\mathbb{R}}_+ \to \tilde{\mathbb{R}}_+$ obtained from f by restricting the domain *and* changing the range is both injective and surjective, so it has an inverse. Its inverse is, of course, what we usually call the *square-root function*.

A useful criterion for showing that a given function f is bijective is the following, whose proof is left to the exercises:

Lemma 2.1. *Let $f : A \to B$. If there are functions $g : B \to A$ and $h : B \to A$ such that $g(f(a)) = a$ for every a in A and $f(h(b)) = b$ for every b in B, then f is bijective and $g = h = f^{-1}$.*

Definition. Let $f : A \to B$. If A_0 is a subset of A, we denote by $f(A_0)$ the set of all images of points of A_0 under the function f; this set is called the ***image*** of A_0 under f. Formally,

$$f(A_0) = \{b \mid b = f(a) \text{ for at least one } a \in A_0\}.$$

On the other hand, if B_0 is a subset of B, we denote by $f^{-1}(B_0)$ the set of all elements of A whose images under f lie in B_0; it is called the ***preimage*** of B_0 under f (or the "counterimage," or the "inverse image," of B_0). Formally,

$$f^{-1}(B_0) = \{a \mid f(a) \in B_0\}.$$

Of course, there may be no points a of A whose images lie in B_0; in that case, $f^{-1}(B_0)$ is empty.

Note that if $f : A \to B$ is bijective and $B_0 \subset B$, we have two meanings for the notation $f^{-1}(B_0)$. It can be taken to denote the *preimage* of B_0 under the function f or to denote the *image* of B_0 under the function $f^{-1} : B \to A$. These two meanings give precisely the same subset of A, however, so there is, in fact, no ambiguity.

Some care is needed if one is to use the f and f^{-1} notation correctly. The operation f^{-1}, for instance, when applied to subsets of B, behaves very nicely; it preserves inclusions, unions, intersections, and differences of sets. We shall use this fact frequently. But the operation f, when applied to subsets of A, preserves only inclusions and unions. See Exercises 2 and 3.

As another situation where care is needed, we note that it is not in general true that $f^{-1}(f(A_0)) = A_0$ and $f(f^{-1}(B_0)) = B_0$. (See the following example.) The relevant rules, which we leave to you to check, are the following: If $f : A \to B$ and if $A_0 \subset A$ and $B_0 \subset B$, then

$$A_0 \subset f^{-1}(f(A_0)) \quad \text{and} \quad f(f^{-1}(B_0)) \subset B_0.$$

The first inclusion is an equality if f is injective, and the second inclusion is an equality if f is surjective.

EXAMPLE 4. Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = 3x^2 + 2$ (Figure 2.3). Let $[a, b]$ denote the closed interval $a \leq x \leq b$. Then

$$f^{-1}(f([0, 1])) = f^{-1}([2, 5]) = [-1, 1], \quad \text{and}$$
$$f(f^{-1}([0, 5])) = f([-1, 1]) = [2, 5].$$

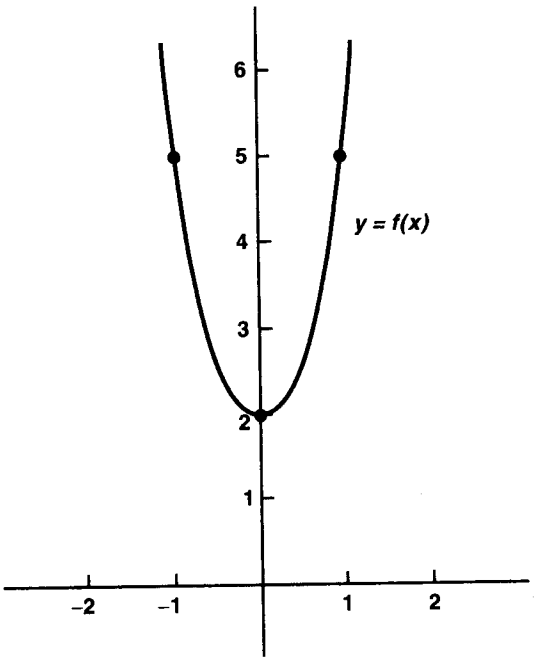

Figure 2.3

Exercises

1. Let $f : A \to B$. Let $A_0 \subset A$ and $B_0 \subset B$.
   (a) Show that $A_0 \subset f^{-1}(f(A_0))$ and that equality holds if f is injective.
   (b) Show that $f(f^{-1}(B_0)) \subset B_0$ and that equality holds if f is surjective.
2. Let $f : A \to B$ and let $A_i \subset A$ and $B_i \subset B$ for $i = 0$ and $i = 1$. Show that f^{-1} preserves inclusions, unions, intersections, and differences of sets:
   (a) $B_0 \subset B_1 \Rightarrow f^{-1}(B_0) \subset f^{-1}(B_1)$.
   (b) $f^{-1}(B_0 \cup B_1) = f^{-1}(B_0) \cup f^{-1}(B_1)$.
   (c) $f^{-1}(B_0 \cap B_1) = f^{-1}(B_0) \cap f^{-1}(B_1)$.
   (d) $f^{-1}(B_0 - B_1) = f^{-1}(B_0) - f^{-1}(B_1)$.
   Show that f preserves inclusions and unions only:
   (e) $A_0 \subset A_1 \Rightarrow f(A_0) \subset f(A_1)$.

(f) $f(A_0 \cup A_1) = f(A_0) \cup f(A_1)$.
(g) $f(A_0 \cap A_1) \subset f(A_0) \cap f(A_1)$; show that equality holds if f is injective.
(h) $f(A_0 - A_1) \supset f(A_0) - f(A_1)$; show that equality holds if f is injective.

3. Show that (b), (c), (f), and (g) of Exercise 2 hold for arbitrary unions and intersections.

4. Let $f : A \to B$ and $g : B \to C$.
(a) If $C_0 \subset C$, show that $(g \circ f)^{-1}(C_0) = f^{-1}(g^{-1}(C_0))$.
(b) If f and g are injective, show that $g \circ f$ is injective.
(c) If $g \circ f$ is injective, what can you say about injectivity of f and g?
(d) If f and g are surjective, show that $g \circ f$ is surjective.
(e) If $g \circ f$ is surjective, what can you say about surjectivity of f and g?
(f) Summarize your answers to (b)–(e) in the form of a theorem.

5. In general, let us denote the ***identity function*** for a set C by i_C. That is, define $i_C : C \to C$ to be the function given by the rule $i_C(x) = x$ for all $x \in C$. Given $f : A \to B$, we say that a function $g : B \to A$ is a ***left inverse*** for f if $g \circ f = i_A$; and we say that $h : B \to A$ is a ***right inverse*** for f if $f \circ h = i_B$.
(a) Show that if f has a left inverse, f is injective; and if f has a right inverse, f is surjective.
(b) Give an example of a function that has a left inverse but no right inverse.
(c) Give an example of a function that has a right inverse but no left inverse.
(d) Can a function have more than one left inverse? More than one right inverse?
(e) Show that if f has both a left inverse g and a right inverse h, then f is bijective and $g = h = f^{-1}$.

6. Let $f : \mathbb{R} \to \mathbb{R}$ be the function $f(x) = x^3 - x$. By restricting the domain and range of f appropriately, obtain from f a bijective function g. Draw the graphs of g and g^{-1}. (There are several possible choices for g.)

§3 Relations

A concept that is, in some ways, more general than that of function is the concept of a *relation*. In this section, we define what mathematicians mean by a relation, and we consider two types of relations that occur with great frequency in mathematics: *equivalence relations* and *order relations*. Order relations will be used throughout the book; equivalence relations will not be used until §22.

Definition. A ***relation*** on a set A is a subset C of the cartesian product $A \times A$.

If C is a relation on A, we use the notation xCy to mean the same thing as $(x, y) \in C$. We read it "x is in the relation C to y."

A rule of assignment r for a function $f : A \to A$ is also a subset of $A \times A$. But it is a subset of a very special kind: namely, one such that each element of A appears as the first coordinate of an element of r exactly once. *Any* subset of $A \times A$ is a relation on A.